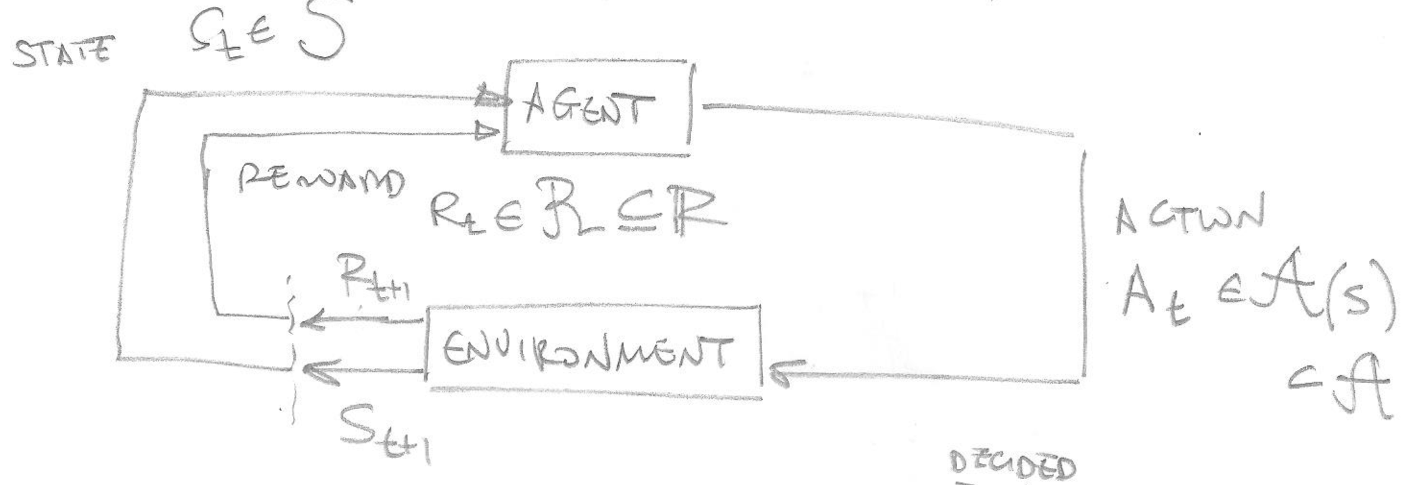
CHAPTER 3: FINITE MARKOV DECISION PROCESSES (MDPs)



TRAJECTORY: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, R_3, \ldots$

DECIDED JOINTLY, SO HAVE COMMON INDEX

MDP IS FINITE, I.E. $|\mathcal{S}|, |\mathcal{R}|, |\mathcal{A}(s)| < \infty$

DYNAMIC FUNCTION

$$p(s', r \mid s, a) \triangleq \Pr\left[S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right]$$

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$$

- IT IS A PMF, THEREFORE:

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1 \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}(s)$$

- THE MDP IS HOMOGENEOUS: LEFT HAND SIDE DOES NOT DEPEND ON TIME t

- THE SYSTEM HAS THE MARKOV PROPERTY: STATISTICS OF S_t, R_t DEPEND <u>ONLY</u> ON S_{t-1} AND A_{t-1}. SO S_{t-1} MIGHT HAVE TO BE LARGE

STATE TRANSITION PROBABILITIES:

$$p(s'|s,a) \triangleq \Pr\left[S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right]$$

$$= \sum_{r \in R} p(s', r \mid s, a)$$

EXPECTED REWARDS

$$r(s,a) \triangleq E\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right] =$$

$$\sum_{r \in R} \sum_{s' \in S} r \, p(s', r \mid s, a) =$$

$$\sum_{r \in R} r \boxed{\sum_{s \in S} p(s, r \mid s, a)} \longrightarrow$$ PROBABILITY WE GET REWARD r

EXPECTED REWARDS FOR (STATE, ACTION, NEXT STATE) 3PLE:

$$r(s,a,s') \triangleq E\left[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'\right]$$

$$= \sum_{r \in R} r \, p(r \mid s, a, s') =$$

$$\sum_{r \in R} r \frac{p(r, s, a, s')}{p(s, a, s')} = \sum_{r \in R} r \frac{p(r, s, a, s') / p(s, a)}{p(s, a, s') / p(s, a)}$$

$\Rightarrow$

$$r(s, a, s') = \sum_{r \in R} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

(SO REWARD STATISTICS DEPEND ON NEXT STATE)

NOTE: WE USED BASIC PROPERTY

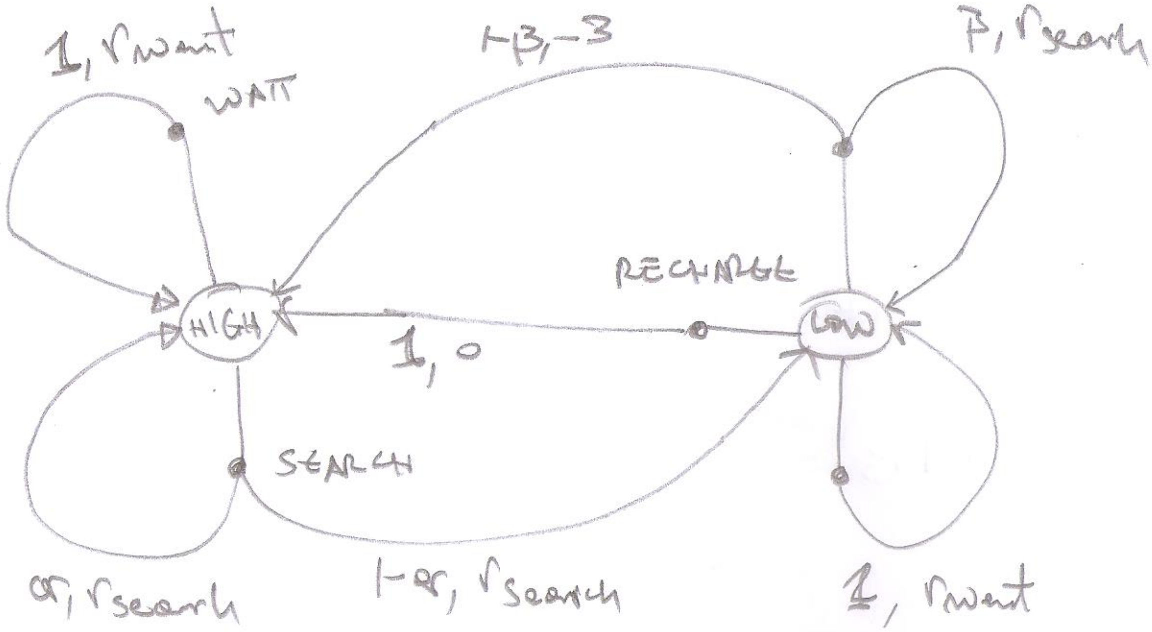
$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

EXAMPLE 3.3. A RECYCLING ROBOT

$S = \{\text{high}, \text{low}\}$ (BATTERY LEVELS)

$A = \{\text{search}, \text{wait}, \text{recharge}\}$

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	r_{search}
-"-	-"-	low	$1-\alpha$	-"-
low	search	high	$1-\beta$	-3
-"-	-"-	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	–
low	wait	high	0	–
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	–



OBSERVE THAT: 1) MARKOVIAN PROPERTY IS NOT JUSTIFIED WELL

2) $r_{search} > r_{wait}$, OTHERWISE SOLUTION IS TRIVIAL

3) WHAT IS OPTIMAL SOLUTION;

3.3 RETURNS AND EPISODES

WHAT DO WE WANT TO MAXIMIZE?

TWO CASES:

1) THERE IS A SPECIAL TERMINAL STATE, WHICH WE ACHIEVE AFTER FINITE R.V. T. THEN AT TIME t WE WANT TO MAXIMIZE THE EXPECTED RETURN:

$$G_t \triangleq R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$$

RETURN, IT IS AN R.V.. THE EVOLUTION OF THE STATES HAPPENS IN EPISODIC TASKS, OR EPISODES

(EXAMPLES: GAMES OF BACKGAMMON)

② CONTINUING TASKS : THERE IS NO TERMINATING STATE AND WE GO FOREVER. AT STEP t, WE WANT TO MAXIMIZE THE EXPECTED DISCOUNTED RETURN:

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

WHERE $\gamma \in [0,1]$ IS THE DISCOUNT RATE

A USEFUL FORMULA FOR THIS CASE:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

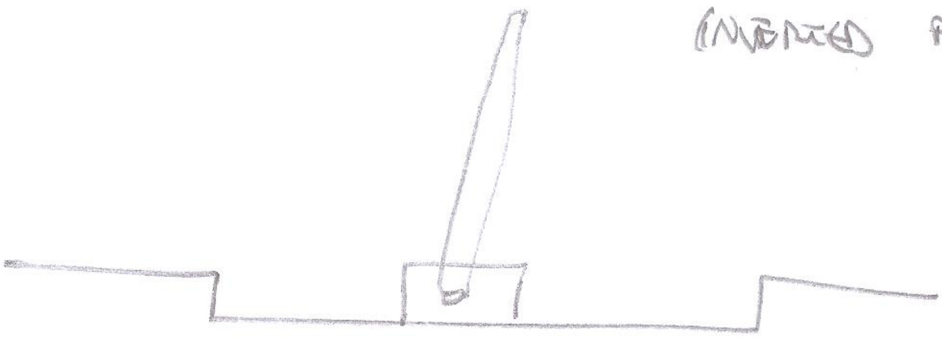
$$= R_{t+1} + \gamma \left[R_{t+2} + \gamma R_{t+3} + \dots \right]$$

$$= R_{t+1} + \gamma G_{t+1} \Rightarrow \boxed{G_t = R_{t+1} + \gamma G_{t+1}}$$

EXAMPLES: CHEMICAL PROCESSES WE DO NOT WANT TO END.

EXAMPLE 3.4 POLE - BALANCING

(INVERTED PENDULUM)



THIS CAN BE MODELED BOTH USING EPISODIC AND CONTINUING TASKS

3.4 UNIFIED NOTATION OF G_t

WE HAVE GIVEN TWO FORMULAS FOR G_t, ONE FOR EPISODIC, ONE FOR CONTINUING TASKS. WE NEED TO HAVE ONLY ONE. WE INTRODUCE TWO CONVENTIONS.

1) FOR EPISODIC TASKS, WE DO NOT INDEX EPISODES SEPARATELY, SO WE ONLY WRITE $A_1, A_2, A_3, \ldots$ AND NOT

$$A_{1,i}, \ A_{2,i}, \ A_{3,i} \ldots$$ FOR EPISODE i

2) WE INTRODUCE ABSORBING STATE

$$S_0 \xrightarrow[R_1]{} S_1 \xrightarrow[R_2]{} S_2 \xrightarrow[R_3]{} \boxed{} \circlearrowleft \quad R_4 = R_5 = \ldots = 0$$

SO, FROM NOW ON, ALWAYS,

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

BUT MAYBE WE NEED $\gamma < 1$

3.5 POLICIES AND VALUE FUNCTIONS

IF AGENT FOLLOWS POLICY π THEN PROBABILITY THAT $A_t = a$ IF $S_t = s$ IS $\underline{\pi(a|s)}$

SO THE POLICY IS WHAT WE WANT TO OPTIMIZE

OBSERVE THAT

$$E\left[R_{t+1} \,|\, S_t = s\right] = E\left[\ E\left[R_{t+1} \,|\, S_t = s, A_t\right]\ \right]$$

$$= \sum_{a} \pi(a|s)\, r(s,a)$$

$$= \sum_{a} \pi(a|s) \cdot \sum_{s',r} r\, p(s',r|s,a)$$

STATE-VALUE FUNCTION FOR POLICY π:

$$v_\pi(s) \triangleq E_\pi\left[G_t \mid S_t = s\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \mid S_t = s\right] \quad \forall s \in \mathcal{S}$$

ACTION-VALUE FUNCTION FOR POLICY π:

$$q_\pi(s,a) \triangleq E_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \mid S_t = s, A_t = a\right]$$

OBSERVE THAT THE TWO ARE CONNECTED:

$$v_\pi(s) = \sum_{a} \pi(a|s)\, q_\pi(s,a) \quad (1)$$

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right] \quad (2)$$

$$\left(= \sum_{s',r} p(s',r|s,a)\, E_\pi\left[G_t \mid S_t = s, A_t = a, S_{t+1} = s', R_{t+1} = r\right]\right)$$

FURTHERMORE, THEY ARE CONNECTED TO THEMSELVES, IN SOME WAY:

$$v_\pi(s) \triangleq E_\pi[G_t \mid S_t = s)$$

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \qquad (3.14)$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma E_\pi\left[G_{t+1} \;\middle|\; \begin{matrix} S_t = s \\ R_{t+1} = r \\ A_t = a \\ S_{t+1} = s' \end{matrix}\right]\right]$$

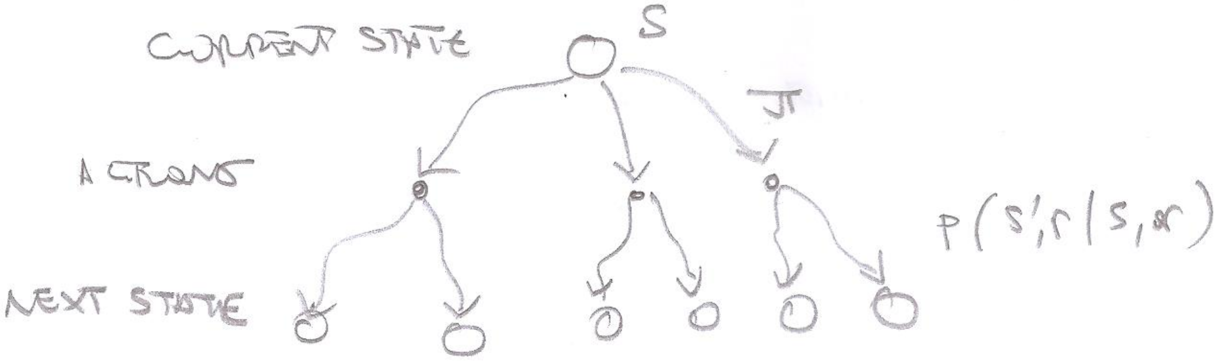
$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma v_\pi(s')\right] \quad \forall s \in \mathcal{S}$$

(ONE OF THEM ANYWAY)

THIS IS THE BELLMAN EQUATION FOR $v_\pi(s)$.

IT FORMS A SYSTEM OF LINEAR EQUATIONS THAT HAS A UNIQUE SOLUTION (USUALLY, THEORY IS TRICKY)

WE CAN CONCEPTUALIZE THIS USING A BACKUP DIAGRAM



THERE IS ALSO A BELLMAN EQUATION FOR

$q_\pi(s,a)$. USING ①, ② :

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \sum_{a'} \pi(a'|s')\, q_\pi(s',a')\right]$$

(THIS IS HARDER TO USE)

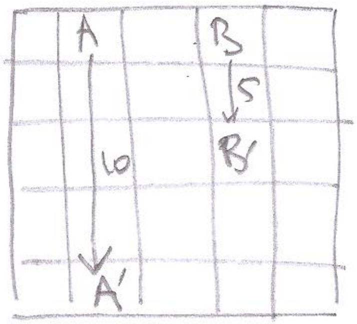
SO, HOW DO WE FIND $v_\pi(s)$?

1) WE CAN DO SIMULATIONS (MONTE CARLO)

2) WE CAN SOLVE A LINEAR SYSTEM

OF COURSE, AIM IS TO FIND <u>OPTIMAL</u> POLICY. WE EXAMINE THIS NEXT.

<u>EXAMPLE 3.5 GRIDWORLD</u>



1) ACTIONS ARE UNIFORM RANDOM WALK EXCEPT FOR STATES A, B

2) REWARDS ARE 0 EXCEPT THOSE SHOWN ON LEFT AND THOSE TAKING AGENT <u>OFF</u> THE GRID, WHICH ARE -1

3) $\gamma = 0.9$

THEN $v_\pi(s)$ IS

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

OBSERVE THAT $8.8 < 10$

$5.3 > 5$

WHY?

NESTED EXPECTATION

LET X, Y TWO R.V.s WITH PMFs

$$p_X(x), \quad p_Y(y), \quad p_{XY}(x,y), \quad p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

WE HAVE

$$E[X] = \sum_x x p_X(x), \qquad E[X \mid Y = y] = \sum_x x p_{X|Y}(x|y)$$

OBSERVE THAT $E[X|Y]$ IS ALSO AN R.V.

THEREFORE, IT HAS AN EXPECTED VALUE!

$$E[E[X|Y]] = \sum_y p_Y(y) \sum_x x p_{X|Y}(x|y) = \circledast$$

(NOTE: • $E[g(Y)] = \sum_y g(y) p_Y(y)$

• $E[g(X,Y)] = \sum_{x,y} g(x,y) p_{XY}(x,y)$)

$$\circledast = \sum_{x,y} x \, p_Y(y) p_{X|Y}(x|y) = \sum_{x,y} x \, p_{XY}(x,y)$$

$$= E(X) \Rightarrow$$

$$E[X] = E[E[X|Y]]$$

NESTED EXPECTATION FORMULA

NICE APPLICATION: MINER'S PROBLEM